Assessment of clinical competence. Written and computer-based assessment

Article in Assessment & Evaluation in Higher Education · September 1987
Impact Factor: 0.84 · DOI: 10.1080/0260293870120307
ASSESSMENT OF CLINICAL COMPETENCE: WRITTEN AND COMPUTER-BASED SIMULATIONS

David B. Swanson, John J. Norcini and Louis J. Grosso
American Board of Internal Medicine

ABSTRACT

Written and computer-based clinical simulations have been used in the health professions to assess aspects of clinical competence for many years. However, this review of the dozens of studies of their psychometric characteristics finds little evidence to justify their continued use. While studies of the fidelity of simulations have demonstrated that examinees feel they are realistic and have good face validity, reliability studies have repeatedly shown that scores are too imprecise for meaningful interpretation, unless impractically large numbers of simulations are included in a test. Validity studies have demonstrated that simulations have the expected relationships with a host of criterion measures, but it appears that similar assessment information can be obtained using clinically-oriented multiple choice questions in much less testing time. Some common methodological weaknesses in study design and analysis are identified, and some research directions are suggested to improve the psychometric characteristics of simulation-based tests.

INTRODUCTION

For over a quarter of a century, written clinical simulations have been used in the health professions to assess aspects of clinical competence. Similarly, computer-based clinical simulations of patient management have been used, usually on an experimental basis, since the late 1960s. Dozens of small and large scale studies have been conducted of the fidelity, scoring, reliability, and validity of these simulations in assessing clinical competence. The purpose of this paper is to synthesise the findings of these research efforts and to summarise the current state of the simulation art. Research studies using written and computer-based clinical simulations are deliberately intermingled, because the two seem to differ very little in their psychometric characteristics.

For purposes of this paper, the term "clinical simulation" (or, equivalently, "patient management problem" or PMP) is used to describe the broad range of linear and branching simulations in which examinees iteratively gather clinical data from and prescribe treatment for a simulated patient. The variations in the format of clinical simulations are seemingly endless: diagnostic management problems\(^{1}\), sequential management problems\(^{2}\), portable patient problem packs\(^{3}\), modified essay questions\(^{4}\), and computerised patient management problems\(^{5}\) are just a few examples. These formats differ in the way they provide information and in what they
require of examinees. For example, some simulations provide all history and physical examination findings in an opening scene or in response to a single examinee action, while others force the examinee to gather this data using a short or long list of options. Some formats use no option lists at all: examinees specify desired data or therapeutic interventions in "short answer" form. Some formats focus on diagnosis and do not include a treatment section, while others concentrate on treatment. More subtly, individual cases differ in the clinical challenges they pose and the knowledge/abilities required for solution (e.g. cases of irritable colon, digoxin toxicity and lupus differ in emphasis on clinical diagnosis, laboratory utilisation, and short/long term treatment).

Clinical simulations have been used for a wide variety of testing purposes (detection/remediation of deficits; grading; licensure/certification) with a variety of examinee populations (medical students, physicians in post-graduate training, practicing physicians). The measurement characteristics of a particular simulation format will vary with the purpose of measurement, the educational level and clinical background of examinees, the format used, and, of course, the quality of the test materials.

Despite the diversity of simulation formats and the variety of testing purposes for which they have been used, some generalisations about their psychometric characteristics emerge clearly from a review of the research literature. These are discussed in four sections, each corresponding to a major issue in assessment with simulations. The first section is concerned with fidelity: the correspondence between the real world and the simulation. The second section reviews studies of various methods for scoring simulations. The third section discusses the reliability (generalisability) of simulation-based scores: the extent to which performance on one clinical simulation predicts performance on other simulations. This topic has received much attention in the medical problem solving literature, where it has been termed "content specificity of problem solving skills". The fourth section reviews evidence for the validity of simulations: the extent to which inferences concerning performance in real clinical situations can be based upon simulation performance. The last section presents a summary of the review, recommendations for use of clinical simulations, and suggestions for further research.

FIDELITY

The original motivation for simulation-based measures was dissatisfaction with assessment techniques available in the 1950s. Many felt that multiple choice questions (MCQs) measured knowledge well, but not the ability to apply knowledge in clinical situations. Oral examinations were thought to assess the latter well, but subjectivity and disagreement in grading performance and making pass/fail decisions were problematic. Ratings of performance based on observation in the clinical environment were also viewed as subjective, and the clinical challenges posed by real patients were unstandardised. Simulation was seen as an objective, standardised technique which could pose important clinical problems in realistic ways.

Because of this emphasis on real-life clinical challenges, several studies have investigated the fidelity of simulations: the extent to which the problems posed and the
actions taken by an examinee are similar to those which occur in a real patient/physician interaction. Three different research paradigms have been used: self-report, chart audit, and simulated patients.

**Self-Report Studies**

Several early PMP research studies, similar to the work of McQuire and Babbot, cite comments from PMP developers and examinees that working through a case "feels like" management of real patients. Studies of computer simulations, have investigated this topic more formally with questionnaires, perhaps because of concern that computer technology would seem artificial and forbidding. These studies have uniformly indicated that both written and computer-based simulations are felt to be realistic and provide a better assessment of clinical medicine than MCQs.\(^7,8,9,10\)

**Audit Studies**

To look more formally at the correspondence between behaviour in real life and on clinical simulations, a number of studies have used retrospective chart audit. Charts document real life activity which can then be compared with simulation performance.

Goran and colleagues\(^11\) contrasted performance on a PMP case of urinary tract infection with similar cases in clinical practice using data abstracted from medical records. Physicians recorded 11% fewer items than they selected on the PMP, with most of the differences on the history and physical examination. Quality of performance of individual physicians was not highly correlated in the two settings.

In a study of sixteen computer simulations presented via the CASE model, the authors reported no differences in the extent to which audit criteria were met in simulated and real cases presenting similar problems.**\(^7\)** Similar findings were obtained in two audit studies comparing performance on CBX computer simulations with similar real life cases.\(^10\)

**Simulated Patient Studies**

Use of chart audit to investigate similarity of performance poses methodological problems. Typically, more data are gathered than recorded in the chart, particularly for history and physical examination. Further, the correspondence between real and simulated cases is difficult to assess, because real patients can differ in ways that justifiably result in data gathering and treatment variation (e.g. co-existing complaints or complications). To control these factors, several studies have used "parallel" written (or computer-based) cases and "live" simulated patients trained to role-play cases.

Feightner and Norman\(^12\) compared performance on two (uncued) CASE computer simulations with analogous simulated patients, using a randomised cross-over design with subjects completing one case in each format. They found that subjects elicited substantially fewer history items (60%), physical examination items (17%), and critical findings (25%) in the computer simulation format. Selection of investigation and management options did not differ in the two formats.
In a similar study comparing performance on four (cued) branching PMPs and analogous simulated patients, Norman and Feighner (13) found that substantially more data were gathered in the PMP format (history — 44% more; physical — 20%; investigations — 140%). In addition, 56% more management options were selected.

A similar study compared the behaviour of pharmacists on four (cued) PMPs with analogous simulated customers. (14) Subjects selected 18% more options on the PMPs, and only low to moderate correlations were observed in the scores obtained on the four matched pairs of cases in the two formats (0.26 to 0.68). The investigators noted that PMPs were better predictors of what pharmacists did not do in practice (options not selected) than of what they did do.

These studies strongly suggest that when presented with options (cueing), subjects tend to make more selections. Moreover, in all of the studies, there were marked individual differences in the options selected on any particular case or section of a case, regardless of the format. This could reflect the redundancy typically present in clinical data: no specific subset of patient data must be gathered to arrive at a diagnosis, because various subsets of patient findings all imply the correct diagnosis and treatment. Given the individual differences in number of options chosen, these findings might also support the notion of a response set or response style. Some subjects have a tendency to choose many options and other subjects have a tendency to choose only a few options. Regardless of the cause, variation in activity level is unrelated to the quality of performance, but it can make scoring difficult.

**Conclusions Regarding Fidelity**

The above studies indicate that simulations provide a reasonably realistic testing vehicle from an examinee’s perspective, but that cueing (and lack of cueing) does affect test performance. In the cued formats, data gathering activities are suggested by the options presented; in the uncued formats, data gathering ideas must be self-generated. As a result, more data is collected in cued simulation formats, and data gathering is generally more thorough, usually leading to higher scores on simulations than in uncued formats or real life. Cueing problems seem to be more severe in the written than computer-based format, because examinees may “read ahead” as well. Options listed in later sections of a case may imply which earlier selections are appropriate. In addition, when working later sections, it is possible to “retrace” one’s steps and select additional earlier options which, in retrospect, are appropriate. Schumacher and colleagues, (15) in a comparison of written and computer-based versions of the same cued PMPs, found markedly better performance on the written versions of the cases with substantially fewer errors of omission. This interpretation is consistent with the results of studies of cued vs uncued PMP formats, (16,17) and work with uncued sequential management problems, (18,2) as well as more general psychological research on the relative difficulty of recall and recognition tasks.

Since data gathering activities with real patients are uncued and open-ended, it would seem that uncued simulation formats are more appropriate. Many of the written
formats can be made less cued through use of long, standardised lists of options in written\(^3^{,19,20}\) and computer-based formats\(^{21}\) or “short answer” requests for data in both formats.\(^4^{,2,18,22,7}\) However, use of uncued formats has psychometric and practical consequences which also must be considered. The vague/ness of the uncued format may increase the importance of “test/wiseness,” which can lead to measurement of abilities not associated with clinical competence; it might also increase the controversy over what constitutes the correct answer(s), leading to less reliable measurement. Similarly, more examinee time is usually required to complete a case in uncued format, and this means that uncued formats may be less reliable per unit of testing time. Additionally, the logistics of test administration become more complex with uncued formats, and machine scoring is difficult or impossible, because of problems associated with the diversity of responses, the equivalence of alternatives, and the (il)legibility of handwriting.

Perhaps, as Martin\(^{18}\) suggests, a reasonable compromise would include history and physical examination sections in an uncued format, because cueing artefacts are clearly serious in these sections, with laboratory and treatment sections in a cued format, because cueing seems to play a less important role in selection of these options.

**SCORING**

*Category Scoring*

**Calculation of Scores.**

Most scoring procedures for written and computer-based simulations involve the classification of options into one of several categories, usually based on the judgment of expert physicians involved in development of the simulations. The categories into which options are classified can vary from “strongly indicated” to “strongly contra-indicated”, with several categories in between. Scoring weights are then assigned to each of the categories; these can vary from a large number of positive points to a large number of negative points, with intermediate values in between. Each option within a category receives the weight associated with the category, and a variety of scores can be calculated by using these weights and the options selected by examinees. For branching simulations, similar weights are also assigned for selection of options that result in transfer from one section to another. Because different examinees respond to different sections and, consequently, have different options (and points) available, additional scoring complexities can occur in branching simulations; these will not be discussed further.

The most common Category Scoring techniques were developed by McGuire and colleagues at the University of Illinois Center for Educational Development. These are termed Proficiency, Efficiency, and Omission Scores.\(^{6,23}\) The Proficiency Score, proposed as a measure of overall skill, is the sum of the points assigned to all selected options divided by the maximum score possible. The Efficiency Score, proposed as a measure of effective management with economy of action, is the proportion of an examinee’s selections which are indicated options. The Errors of Omission Score (also
termed the Select or Thoroughness Score when a different scaling is used) reflects the proportion of beneficial options that were not selected. The Errors of Commission Score (also termed the Non-Select or Omit Score) reflects the proportion of contraindicated options selected. Usually, these scores are calculated by aggregating across all sections of a simulation, so that each reflects a somewhat different dimension of overall performance on a case. Specific formulas vary, depending upon investigators' preferences and format used.\(^{(24)}\)

A variant on these methods subdivides cases into naturally occurring clinical tasks, and assigns a score to each, e.g. History-Taking, Physical Examination, Laboratory, and Treatment Scores. Sometimes, these scores are further aggregated into a Data-Gathering Score and a Decision-Making Score. Generally, these scores are calculated much like the Proficiency Score, with the appropriate section(s) contributing to score calculation (e.g. history sections contribute to the History-Taking Score). Computational details vary from study to study.

Studies of Category Scoring.
Bligh\(^{(25)}\) compared nine scoring systems which varied in the number of categories used (from three to eleven categories) and the relative magnitude of the weights assigned to each (−16 to +16 vs −3 to +3 vs −1 to +1, etc.). He found that scores were highly correlated (at least 0.9 and usually higher), regardless of the number of categories used and points awarded, as long as a reasonable proportion of options (40%) had non-zero weights (were classified as either indicated or contraindicated). Donnelly, in a smaller study, compared two scoring schemes which varied along similar dimensions with essentially identical results.\(^{(26)}\)

Nocini and colleagues\(^{(24)}\) compared a variety of scoring systems with the Proficiency Score used by the American Board of Internal Medicine in its Certifying Examinations. They found no differences between scoring systems in ability to predict two external criterion measures. Using regression techniques,\(^{(27)}\) they demonstrated that choice of weights was unimportant, as long as indicated options had positive weights and contraindicated options had negative weights. As in the Bligh study,\(^{(25)}\) high correlations between scoring schemes were observed.

In contrast, Donelly and colleagues\(^{(28)}\) found that changes in scoring weights did have an impact. Working with ten linear PMPs, scores were initially calculated using weights of 0 and +1 for History and Physical Examination Scores and weights of −1, 0 and +1 for Laboratory Scores. Noting that History and Physical Examination Scores were usually more reliable (more highly correlated with analogous scores on other cases), the investigators merged the −1 and 0 categories together and used a uniform weight of 0. This change increased the reliability of the Laboratory Score (coefficient alpha based upon Laboratory Scores for the ten cases) from 0.77 to 0.87. Unfortunately, the investigators reported no additional information concerning the modified Laboratory Scores, which may simply have measured amount of laboratory utilisation, rather than quality. This effect has been observed as an artifact of PMP cueing in other studies.\(^{(29,24)}\)
Aggregate Scoring

Score Calculation
The Aggregate Score technique, developed by Norman, uses an empirical, rather than the judgmental approach to determination of the scoring weights assigned to each option. In this scheme, a criterion group, usually composed of expert physicians, is asked to work through a simulation as an examinee would. Each option is assigned a weight equal to the proportion of the expert group who selected it. An examinee's score is calculated as the sum of the weights of the selected options divided by the maximum possible score for that number of options. Scores, thus, reflect level of agreement with expert selections, adjusted for number of selections made. There are a number of variations on this approach to scoring, with most altering the divisor in score calculation to prevent examinees from being rewarded for picking inappropriately small numbers of options. This method is most often applied to uncued simulations, because it defines a set of acceptable responses at the same time that it weights them. Nonetheless, it is also applicable to cued formats.

Studies of Aggregate Scoring
Only a small number of studies have used Aggregate Scoring methods for written or computer-based simulations. Investigators from the American Board of Internal Medicine used Aggregate Scoring for 16 linear PMPs from the 1980 Certifying Examination in Internal Medicine, as well as several of the more popular Category Scoring methods. Results indicated relatively large correlations among all scoring methods and no substantial differences in reliability, validity, or incremental validity (correlations with external criteria over and above MCQs).

In another study, Norcini compared the weights derived through Aggregate Scoring with those assigned by test developers using Category Scoring methods. Virtually all options identified as “indicated” (receiving positive scoring weights) in Category Scoring were also selected by a majority of the criterion group used in Aggregate Scoring; the reverse was also true. Hence, it is not surprising that there is good correspondence between Category Scoring and Aggregate Scoring methods.

Other Scoring Methods

Marshall’s Scoring Method
Marshall reported a modification of traditional Category Scoring techniques that appears promising, but requires further investigation. Marshall reasoned that the usual Proficiency/Efficiency Scores penalised the efficient problem solver who gathered only the minimum data required to make accurate clinical decisions. Consequently, his scoring system subtracted a small number of points for selecting options that were neither indicated nor contraindicated (these are normally given a weight of zero) to penalise excessive data gathering. It also set a maximum number of points which could be earned in each section, with the maximum being less than the sum of the positive points available in the section. This avoided rewarding examinees for gathering excessive, redundant clinical data, while allowing natural variation in the specific data.
which individual examinees did obtain. In a small scale study of the scoring system using long branching PMPs in a relatively uncued format, these manipulations resulted in the highest overall scores for efficient problem solvers (all else being equal) and also increased the correlation between cases from about 0.3 to 0.5, thereby improving total test reliability. No other investigators have studied this scoring system, however.

Boolean Scoring Methods
Several studies have explored the use of "Boolean" scoring methods, in which logical combinations of options are scored rather than individual options. Webster\(^{31}\) explored two such methods in conjunction with linear PMPs appearing on the Certifying Examination of the American Board of Internal Medicine: "Absolute Scoring" and "Goal-Oriented Scoring". In the former, the investigators identified those options which would expose the patient to serious risk of harm. A score of one was assigned to a PMP if an examinee selected all essential options and avoided all harmful options; a score of zero was received for the case otherwise. In the Goal-Oriented system, scoring is anchored to the key clinical goals in a case (e.g. in a patient with anemia, these might include establishing the cause of anemia, its severity, and instituting treatment for the anemia and the underlying cause). Rules were then developed for scoring options related to the same clinical goal; scores for each goal were then combined into an overall case score. Comparison of Absolute and Goal-Oriented Scores with several Category and Aggregate Scoring alternatives yielded no differences in reliability or validity, however, except that Absolute Scores were less reliable, as one might expect, given the loss of measurement information in using zero/one scores only. Slotnick\(^{33}\) has also explored some other Boolean scoring systems.

Pathway Scoring Methods
Another scoring approach that has been used with computer-based clinical simulations focuses on key clinical decisions and their impact on the "patient's condition". Two forms of these "Pathway Scores" were used in a field test of the CBX computer-based simulation: Decision-Point and Endpoint Scores.\(^{10}\) For the former, three to five key clinical decisions were identified in each case, and points were awarded for correct decisions and subtracted for wrong decisions. For the latter, the patient's condition at the end of the case (which was primarily a function of these key decisions) was classified as "optimal", "satisfactory", or "unsatisfactory", and received an associated point score. These scores correlated much better than more traditional scoring methods with expert judgements of examinee performance on the associated cases. Unfortunately, these scores were also the least reliable indicators of an examinee's performance over cases. No other studies have explored analogous Pathway Scores.

Conclusions Regarding Scoring
It appears that changes in the number of categories into which options are classified and moderate changes in the weights assigned to those categories make little difference in the resulting scores. Similarly, differences between Aggregate Scoring and Category Scoring systems are small. Correlations between these approaches are so high that the
reliability and validity of scores are virtually unaffected. This finding parallels results of measurement research on use of weighted vs unweighted composite scores\(^{34}\) as well as research on the accuracy of prediction using linear composites with different (regression) weights\(^{27}\).

Marshall's work\(^{19}\) involved more complex changes to traditional scoring methods than simple changes in weights. It is unclear without further research if the improved reliabilities affect validity; this may depend upon the characteristics of the specific cases used and the abilities which a test is intended to assess. Similarly, more research is needed on the psychometric characteristics of Boolean and Pathway Scoring methods.

**RELIABILITY**

Reliability refers to the consistency or reproducibility of scores — the extent to which variation in a set of grades represents systematic differences among individuals, rather than other sources of (error) variation\(^{35}\). Virtually all of the studies of clinical simulations have used classical test theory to investigate reliability. Unfortunately, because classical theory does not provide good tools for conceptualising and estimating reliability, much of the work on the reliability of clinical simulations has been technically inadequate. Classical theory encourages the estimation of reliability using a one-source-of-error-variance-at-a-time approach, resulting in a plethora of different interrater, intercase and intracase reliability coefficients. These are often overestimates of actual reliability, because, in reality, scores are simultaneously influenced by all sources of error variance. Generalisability theory\(^{36}\) provides an integrated framework and appropriate techniques for estimating reliability (generalisability) under these circumstances, as well as a better conceptualisation of reliability than relates it to the intended use of the scores. The concepts of generalisability theory guided the review of these studies and have strongly affected the interpretation of results presented in this section.

Studies of the reliability of clinical simulations have investigated the magnitude of three major sources of error variance: (1) interrater variation in the development of the answer key, (2) intercase variability in performance, and (3) intracase variation. Reliability studies investigating each of these sources of error variance are reviewed below.

**Interrater or “Answer Key” Reliability**

There has been serious concern that scoring techniques for clinical simulations do not adequately reward/penalise different approaches to patient workup and treatment\(^{19,37}\). This concern is generated in part by the substantial variation, even among skilled physicians, in performance on any particular real or simulated case\(^{38,39}\). Such variation reflects the redundancy of clinical findings as well as disagreement over what constitutes optimal patient management. In scoring simulations this variation results in lack of agreement on the correct answers.
Several studies have investigated expert agreement in scoring of clinical simulations. Sedlacek and Nattress\textsuperscript{(40)} reported results of a study of interrater agreement of seven judges in constructing a key for one linear PMP taken from the 1969 Orthopedics Certification Examination. There was "a high level of interjudge agreement", with the median correlation between pairs of judges equal to 0.77. A majority of the seven judges agreed on 61\% of the items; all seven agreed exactly on 17\% of the items. The nature and magnitude of disagreements was not discussed. Mazzuca and Cohen\textsuperscript{(37)} investigated interrater agreement on two branching PMPs concerning management of diabetes. They found that the agreement level (percentage of items rated identically) among all possible pairs of raters ranged from 0.31 to 0.73, with a median of 0.43, depending upon the case and section of the case (chance level of agreement = 0.20). Despite the relatively low level of agreement, the average rating given to an option predicted the percentage of endocrine fellows selecting an option fairly accurately (median correlation between average rating and percentage selecting was 0.82 with a range of 0.60 to 0.97).

A more recent study\textsuperscript{(32)} investigated the impact of variability in answer key construction on total errors of measurement in a set of 12 PMPs and compared several methods for minimizing this effect. Results indicated that several scorers were needed in order to reduce measurement error.

Taken as a whole, these studies indicate that the judgment of several experts is required to develop an answer key regardless of the scoring method used. The diversity of opinion among experts precludes key development by a single individual.

\textit{Intercase Reliability}

The purpose of administering clinical simulations goes beyond assessing performance on the particular cases used. Simulations are intended to provide a general assessment of examinees' ability to apply knowledge in some clinical domain, e.g., problems that present with coma, problems of diabetic management, endocrine problems, common internal medicine problems, etc. The particular cases used "stand for" this domain, much as the items on a multiple-choice test are viewed as a sample of items from a broader domain. A set of simulations will have a different reliability, depending upon the breadth of the clinical domain to which the overall test score is expected to generalize and the relationship (representativeness) between the sample of simulations and the full domain. A small number of cases involving diabetes may predict how an examinee would score on other diabetes simulations fairly well; the same cases might predict performance in the broader domain of internal medicine much less accurately.

Unfortunately, most research on the reliability of clinical simulations has not clearly identified the domain to which test scores should generalize, and explicit definitions of domains have not guided case and test construction.\textsuperscript{(41,42,43)} However, many have implicitly viewed problem-solving ability as the domain or general trait to which they wish to generalize from scores on simulations. According to this view, different clinical simulations provide equally good contexts for the display and measurement of problem-solving skills. This approach to assessment of problem solving is congruent with the
unstructured "master-apprentice" approach to clinical training used in most medical schools, but is not consistent with the results of research on medical problem solving.\(^{(39)}\)

Research on medical problem solving has indicated that clinical problem solving cannot be viewed as a general trait displayed with equal skill over a broad range of clinical areas; problem solving skills seem to be content-specific. Physicians have profiles of competence which vary from specialty to specialty, and, indeed, from case to case within a specialty area.\(^{(39,38,44)}\) These profiles of competence translate into large differences in quality of performance from case to case in medical problem solving research. The psychometric implications of this for simulations are low correlations between scores on different cases and poor reliability for composite test scores that are based upon a small number of clinical simulations.

Although it is typical to compare the intercase reliabilities of different studies, the primary statistic for this review will be the average intercase correlation — the average correlation between analogous scores on different cases (i.e., history score on two cases). Intercase reliability varies with text length (the number of cases), which varies across the studies reviewed. Consequently, use of intercase reliability would make it difficult to discern patterns of findings. The average intercase correlation, on the other hand, is unaffected by test length and provides a comparable statistic across studies that is easily translated into a reliability coefficient through the Spearman Brown formula.\(^{(34)}\)

Although a large number of studies have used tests consisting of two or more clinical simulations, relatively few have reported intercase correlations. This section reviews studies that have reported the correlations directly or have included sufficient information (interease reliability) for the mean intercase correlation to be computed.

The first study in this group was discussed previously in the section on scoring techniques. Marshall\(^{(19)}\) correlated overall performance on two (relatively) uncued PMPs used in the Royal Australian College of General Practice certifying exam in 1974 through 1976 (sample size varied from 113 to 188). These correlations were 0.45, 0.47 and 0.61 respectively, when the PMPs were scored using Marshall's revised technique. These values were substantially higher than those found prior to 1974 when standard PMP scoring techniques were used.

Berner and colleagues\(^{(45)}\) administered five uncued sequential management problems to 190 second year medical students, generating four scores for problem list formulation, selection of correct diagnostic procedures, selection of incorrect diagnostic procedures, and final diagnosis. Intercase correlations between analogous scores on different cases ranged from 0.00 to 0.47; median correlations ranged from 0.15 (problem list score) to 0.31 (final diagnosis).

A similar study,\(^{(46)}\) using 69 third year students, four sequential management problems (cued in some sections, uncued in others), and five scores per case, found that analogous scores on different cases correlated from —0.17 to 0.31, with median intercase correlations ranging from 0.02 to 0.09, depending upon the score/ability involved.

Donnelly and colleagues\(^{(47)}\) administered ten branching PMPs to 162 third year medical students. Mean intercase correlations ranged from 0.02 (treatment) to 0.40
(history), with physical examination and laboratory scores showing moderate intercase correlations (0.23), and radiographic and diagnosis sections having low intercase correlations (both averaged approximately 0.07).

Wolf and colleagues\(^{48,49}\) administered four linear PMPs and eleven similar PMPs as a pre-test and post-test (respectively) to 175 medical students before and after a four-week "Interphase Programme" designed to prepare the students for clinical clerkships. All cases included cue sections for history, physical examination, diagnostic studies, and therapeutic procedures, plus a differential diagnosis section in short answer format; Proficiency Scoring was used to calculate a composite score for each case. Correlations between case composite scores averaged 0.21 for both the pre-test and the post-test.

Donnelly and colleagues\(^{28}\) report somewhat stronger relationships between scores on different cases. Using 10 linear uncued PMPs, all in the area of chronic obstructive pulmonary disease, and 46 subjects ranging from fourth year medical students to pulmonary specialists, intercase correlations in the 0.3 to 0.8 (average of 0.58) range were obtained between history sections on different cases; correlations between physical examination scores were in the 0.2 to 0.7 range (average of 0.5). Laboratory and treatment scores showed lower intercase correlations (averages of 0.24 and 0.19, respectively, in one scoring system and 0.4 for both using the second scoring system (discussed in the section on scoring above). The extreme range of ability in the study sample and the very limited content domain probably explain the higher intercase correlations.

Robinson and Dinham\(^{50}\) in a small scale study involving 41 medical students and two sets of three diagnostic management problems used as pre- and post-tests in a pediatrics clerkship, reported average intercase correlations ranging from 0.03 (diagnosis scores) to 0.67 (proficiency scores, probably emphasizing history and physical examination).

Norcini and colleagues\(^{51,52}\) in psychometric analyses of the 1980, 1981 and 1982 certifying examinations in internal medicine (4000 subjects each year), reported average intercase correlations of 0.14 to 0.18 among the twelve to sixteen cued linear PMPs (focused on lab utilisation and treatment) used in each year. Comparable low intercase correlations have been observed on all of the ABIM subspecialty certification examinations using PMPs.\(^{53}\) The American Board of Pediatrics reports similar low intercase correlations for the cued, linear PMPs formerly used on their certifying examination in general pediatrics.\(^{54}\)

Psychometric studies of computer-based clinical simulations have reported intercase correlations very similar to those with written simulations. Skakun and co-investigators\(^{55}\) reported intercase correlations ranging from 0.12 to 0.49 (median of 0.24) among four computerised cued PMPs (three linear, one branching), based upon a sample of 71 subjects taking the PMPs as part of the Canadian certification examination in pediatrics. Norcini and co-investigators,\(^{9}\) using the uncued CASE system, administered several different sets of three clinical simulations to groups of 49 to 95 practicing internists. Intercase correlations averaged about 0.3, although they were somewhat higher for history and physical examination skills and somewhat lower for
laboratory utilisation, diagnosis and therapy. In a study involving 153 interns, third year internal medicine residents, and practicing physicians, the six clinical simulations focusing on acute, hospitalised management were administered using the uncued CBX model.\(^{10}\) Average intercase correlations ranged from 0.39 (diagnosis and therapy scores analogous to PMP proficiency scores on those sections) to 0.07 (patient outcome score).

Most of the studies reviewed above reported mean intercase correlations in the 0.1 to 0.4 range. These correlations indicate that performance on one case is a poor predictor of performance on other cases: content specificity consistent with research or medical problem solving. In order to compensate for the low intercase correlations, it is necessary to use a large number of cases to achieve acceptable levels of intercase reliability, exactly as one uses large numbers of MCQ items to develop a reliable total test score. If intercase correlations average 0.1, and a total test reliability of 0.8 is desired (a minimum value for assessment of individual performance), then 30 to 40 cases are required; if intercase correlations average 0.4, approximately six cases are needed to achieve a reliability of 0.8.

Intercase correlations for history and physical examination sections tend to be somewhat higher than other intercase correlations. This may mean that clinical simulations are sufficiently reliable for assessment of skills in these areas. This result could also be due to scoring artifacts related to thoroughness induced through cueing or to consistent differences in examinee perceptions of the “test task” posed by clinical simulations.\(^{18,56}\) Further research on what these scores measure, with careful attention to effects of cueing and test-taking strategy, seems merited. The low intercase correlations observed between treatment sections of different cases suggest that simulations cannot be used effectively to assess skills in this area without prolonged testing time. Additional, larger scale work, using the Marshall\(^{19}\) approach to scoring also seems desirable, since this work represents the sole demonstration of reasonably high intercase correlations for overall case scores.

**Intracase Reliability**

A number of prior studies have reported reliabilities for individual cases in the 0.8 range.\(^{1,6}\) These reliabilities appear to have been calculated by applying coefficient alpha (or, equivalently, KR-20) to the items within a case(s), with items keyed correct if they are appropriately chosen (or appropriately avoided) and incorrect otherwise. Schumacher and colleagues\(^{15}\) report use of a similar approach with three cases simulated by the INDEX computer model, with items from different cases combined into a single subjects-by-items data matrix. Use of this approach has also been reported in calculating reliabilities for the PMPs on the NBME Part III Examination\(^{57}\) and the ABIM Certifying Examination.\(^{58}\) Invariably, this technique results in high reliabilities, regardless of simulation format.

The technique used in these studies to estimate reliability assumes that items are randomly sampled from the domain of interest and are independent (performance on one item does not influence performance on other items). These assumptions are met (more or less) by MCQs and by scores on simulations at the case level. Items within a case, however, are in no sense randomly sampled from a domain: they are
specifically chosen because of their relevance to the simulation. Further, items within a simulation are designed to be dependent: they are deliberately included so that the (correct) selection of a particular item influences the probability of selection of other items. This is, in fact, the essence of problem solving ability which simulations are presumed to measure.

As mentioned above, reliability is a function of the magnitude of the correlation between items. Violations of sampling and independence assumptions spuriously increase these correlations, thus resulting in seriously biased and inflated reliability coefficients. When items from several cases are combined, overestimation decreases somewhat, because some pairs of items will be independent — they come from different cases. Unless the number of cases is very large and the number of options within a case very small, the magnitude of the effect on reliability will be quite large. The correct estimation procedure is the one implied in the discussion of intercase reliability above. Individual case composite scores (or analogous component scores from different cases) should be treated as if they were items: coefficient alpha (or another appropriately chosen intraclass correlation\(^{(36)}\)) should be computed from a subjects-by-case data matrix. Cases are independent and the natural unit for "random" sampling from a domain, so assumptions are satisfied, and the sampling method is consistent with the estimation procedure. This technique yields a reliability estimate that can be interpreted as a correlation coefficient between the observed score and the true score for the domain of interest. If the reliability estimate is low, it indicates that too few cases were used to arrive at an accurate estimate of performance in the target domain, exactly as on a multiple choice test.

**Conclusions Regarding Reliability**

The intercase source of error variation is clearly very large and seriously affects the reliability of clinical simulations. Interrater agreement and choice of scoring procedures may also lower reliability. Most of the reliability estimates from the studies discussed in this section were derived using classical true score theory techniques. The actual reliabilities are somewhat lower, because these techniques do not reflect the fact that the sources of error variation are additive. The low intercase reliabilities will become even lower when interrater variability is included in the error variance. At a minimum, eight to ten cases are probably required to achieve adequate precision for testing purposes, with a committee of experts developing the answer. In many situations, several dozen cases (and one or two days of testing time) are required for accurate measurement. Shorter tests involving a few cases are not sufficiently reliable for individual testing and simply should not be used.

These results are in sharp contrast to the testing time required to achieve acceptable levels of reliability with multiple choice tests, where no more than two or three hours of testing time (about one hundred items) are typically necessary to achieve a reliability greater than 0.8. Low intercase correlations dictate broad sampling, and multiple choice tests can sample clinical areas more broadly in a fixed amount of testing time.\(^{(51,52)}\) Of course, simulations may measure different abilities than multiple choice tests but this is a validity issue, the topic of the next section.

233
VALIDITY

Validity refers to "the accuracy of a prediction or inference made from a test source".\(^{59}\) It is not a property of the test itself, but of interpretations based upon the test scores. Thus, the same test has many validities, depending upon how it is used (e.g. for making grading vs licensing decisions; for inferring problem solving ability in narrower vs broader domains; for measuring data gathering vs decision making skills; for assessing the ability of medical students vs residents vs practitioners). This implies that (1) clinical simulations may be valid for some purposes but not for others; (2) determination of validity is difficult because of the diversity of the formats, abilities, purposes, and interpretations intertwined in use of simulations, and (3) no single study can provide definitive evidence for or against validity.

For purposes of this paper, validity studies will be divided into two categories. Internal validity studies, the first category, include a discussion of content validity and investigations of the dimensionality of written and computer-based simulations. Criterion validity studies, the second category, include comparisons of simulation scores with external criteria such as ratings and scores on other measures.

Reliability and sample size are two factors that affect the quality of validity studies. Because the reliability of simulation-based tests tends to be very low, studies were not included in this review unless substantial numbers of cases were used. In addition, because validity tends to be judged by the relative magnitude of correlation coefficients, studies with small numbers of subjects and unstable correlations are not reviewed.

*Internal Validity Studies*

Content Validity

Most investigators have assumed that simulations possess a high degree of content validity because of the high "face validity" of simulations. As reviewed in the section on fidelity, examinees indicate that working through a simulation "feels like" real clinical activity. Although the evidence concerning the fidelity of simulations is not overwhelmingly positive, a number of investigators have asserted that the similarity of the problem solving tasks in patient care and on clinical simulations is sufficiently great that simulations are content valid.\(^{64,44}\)

Evidence concerning similarity of behaviour on similar real and simulated cases is, indeed, relevant to an assessment of content validity. However, equating content validity and fidelity is almost certainly incorrect for several reasons. First, even if behaviour on real and simulated cases were identical, a valid method for scoring behaviour would still be necessary. It is probably more important that "scores" on real and simulated cases correspond, and there is little evidence that this is true. Second, for a test to be content valid, it is important that the simulated cases be sampled from the real world clinical domain in such a way that they are representative and test important skills.\(^{60}\) A good deal of judgment enters into the choice of simulation material as a consequence. Poor selections can easily be made, just as they
could be if one were assessing performance directly in real life. Cases pose different clinical challenges and vary in the importance of the abilities they assess. Last, the discouraging findings reviewed in the subsection on intercase reliability indicate that a large sample of cases are necessary to develop an accurate estimate of ability. Tests using small numbers of cases cannot be content valid, because they do not sample clinical situations sufficiently for accurate estimation of scores.

Dimensionality of Simulations
A number of studies have investigated intratest relationships between different scores derived from performance on simulations, using psychological theories of medical problem solving to guide interpretation. Donnelly and colleagues\(^{(47)}\) factor analysed the seven scores produced by each of ten branching PMPs on a test administered to 162 third year medical students. For most of the cases, two orthogonal factors were obtained, one with high loadings for some combination of history, physical examination, laboratory and radiograph scores (data gathering factor) and one with high loadings for some combination of diagnosis and treatment scores (decision-making). Factor analysis of these seven scores averaged across cases confirmed this structure.

In a similar study of two hundred junior medical students and 24 branching PMPs, Juul, Noe and Nurenberg\(^{(61)}\) found a similar data-gathering/decision-making factor solution using an oblique rotation (factors correlated 0.4). This structure was replicated on a study of the same students a year later with a new test and on the next cohort of junior students.

Berner and co-investigators\(^{(45)}\) obtained similar results in a study of 190 second year medical students using five sequential management problems. Twenty variables (four scores per case times five cases) were subjected to an alpha factor analysis, which yielded four factors corresponding to the four scores (diagnosis, ordering correct diagnostic procedures, initial problem list, and incorrect diagnostic procedures). It appeared that the sequential management problem format decreased correlations between scores on a given case by reducing the dependence between the sections through feedback on correct selections. This clarified the intercase relationships between the scores.

Skakun,\(^{(62)}\) however, using similar analytic procedures on five linear PMPs administered to 103 final year medical trainees, was unable to produce an interpretable factor structure. Twenty scores (four scores per case times five cases) produced nine factors with no clear pattern within or between cases. However, when the four scores were averaged across problems and factor analysed, a two factor solution was obtained: data gathering (history, physical, laboratory scores) and management.

Harasym and colleagues,\(^{(22)}\) in a study using seventy-one second year (three year curriculum) medical students and three cases in a “multiple stations” format (a mixture of simulated patients, multiple choice, and short-answer questions aimed at different facets of clinical competence), found general “data gathering” and “hypothesis generation” factors (principle components analysis, Harris-Kaiser transformation procedure for components with eigenvalues greater than one, oblique rotation; factors correlated 0.2 to 0.3) to underlie problem solving performance, as well as a “medical knowledge” factor that was case-specific. No treatment component was included in the cases, so the results are quite close to the previous studies.
Donnelly and colleagues\(^{(28)}\) using ten linear PMPs involving chronic obstructive pulmonary disease and 41 subjects of diverse training/experience (fourth year students through pulmonary specialists), factor analysed the ten case scores for history, physical, lab and treatment separately with one-, two-, two- and one-factor solutions, respectively. These findings were interpreted as strong evidence for generality of problem solving skills.

Conclusions Regarding Internal Validity
Taken as a whole, these research results demonstrate little more than considerable study to study variability. A number of investigators have found that two abilities, data-gathering and decision-making, underlie simulation performance. However, each study uses different analytic techniques, usually some type of factor analysis. It is unclear if factor analysis is a strong “discovery procedure” for investigating underlying traits, particularly when sample sizes are small and intercase correlations are low, both of which affect stability of the factor structure. Intuitively, the oblique factor analysis techniques used in several of the studies make good sense, since one would expect data-gathering and decision-making ability to be correlated. Studies like that of Juul\(^{(61)}\) with relatively large sample sizes, long tests, and several replications, are particularly convincing.

Criterion Validity Studies
Relationship to Multiple Choice Tests
Since simulations were developed, in part, because of dissatisfaction with the ability of MCQs to assess problem solving, a number of investigators have studied the relationship between scores on simulations and multiple choice tests. In most of these studies, a “low to moderate” correlation (0.2 to 0.5) was obtained.\(^{(63,10,6,64,48,49,55)}\) Often, this has been interpreted as indirect evidence that simulations are measuring “something different”: problem-solving ability not tapped by multiple choice tests. This is, indeed, a possible explanation of the results.

However, an equally plausible explanation is that the low observed correlations are due to reliability-based attenuation in the magnitude of the coefficients. In one of two studies investigating this possibility directly, Norcini and colleagues\(^{(51,52)}\) report observed correlations of 0.75, 0.71 and 0.76 between the (eight hour) multiple choice and (4.5 to 6.0 hour) linear PMP sections of the 1980, 1981 and 1982 certifying examination in internal medicine (4000 examinees per year). Multiple choice test reliabilities in those three years averaged 0.92 (coefficient alpha); PMP reliabilities averaged 0.73 (intercase reliability). Using the Spearman Brown correction for attenuation,\(^{(34)}\) true correlations between the multiple choice and PMP subtests were calculated to be 0.92, 0.86, and 0.97. These results strongly suggest that MCQ and PMP tests measure similar (or very highly correlated) traits. A similar study with PMPs and MCQs from the subspeciality examination in cardiovascular disease produced similar results.\(^{(65)}\) Of course, different results might be obtained with different simulation formats: the linear PMPs used in these studies focus on laboratory diagnosis and management, so it is unlikely that a strong general data-gathering factor is present in the PMP scores.
In another study, Norcini and co-investigators\(^9\) compared performance on the CASE computer simulation with performance on PMPs and MCQs. Corrected for attenuation, the correlations between the PMPs and MCQs were 0.80 while the corrected correlations between CASE and the PMPs and MCQs ranged from 0.58 to 0.62. The corrected correlations between the MCQs/PMPs and CASE component scores for history and physical Examination, however, were relatively low (0.29 to 0.39); the corrected correlations between the MCQs/PMPs and CASE component score for laboratory utilization were moderate (0.59 to 0.66); and the corrected correlations between the MCQs/PMPs and CASE component scores for diagnosis and management were very high (0.64 to 1.0). These results suggest the presence of a strong data-gathering factor in the computer simulation that is not assessed in MCQs.

Overall, it is unclear how much the low to moderate correlations reported in the literature are attenuated by low reliability. Investigators rarely report the reliability of the multiple choice and simulation tests used in studies. There was a fairly clear trend in the literature for studies using shorter simulation tests (few cases) and short, "home-grown" multiple choice tests to report lower correlations.

**Relationship with Group Performance**

A number of studies have compared the relative performance on simulations of medical students, residents, and experienced practitioners with the expectation that those with more experience would perform better. Almost invariably, medical students at more advanced levels of training perform better than medical students at an earlier stage, and residents perform at a still higher level, with more advanced residents performing best.\(^{6,58,64,15,66,10,56,24}\)

These results are independent of the simulation format used. The findings are not too surprising: given a year or two of additional education and clinical experience, it would be more surprising if the groups did not differ. A carefully constructed ten-item multiple choice test could probably demonstrate differences between groups, solely on the basis of knowledge gains.

Studies comparing the performance of practicing physicians with medical students and residents are more interesting from a theoretical perspective. Practicing physicians generally perform more poorly on knowledge-oriented tests, reflecting both changes in medicine since training and natural decay of non-essential knowledge. Clinical experience should increase problem solving skills, so simulations may show better performance for practitioners.

Page and Fielding\(^{14}\) demonstrated that practicing pharmacists perform better on written simulations. Norman and colleagues\(^{67}\) obtained similar results in a comparison of medical students and practitioners. Martin\(^{18}\) in contrast, showed that practicing physicians performed worse than medical students on uncued sequential management problems and on laboratory, therapy, and management sections of cued PMPs; they performed better than students on cued history and physical examination sections. In a small study involving a single case, Newble, Hoare and Baxter\(^{17}\) found that student performance improved with increasing training on both written and oral simulations, but practitioners performed better only on the oral version. They attributed the results for practitioners to the need for less data to reach the correct decisions; practitioners would receive lower scores as a consequence by being more efficient.
Study of the CBX computer model showed that practitioners perform more poorly than third-year residents and more like interns on all scoring indices but unnecessary cost and risk.\(^{(10)}\) This was interpreted as a promising result, since avoidance of cost and risk are problem-solving skills that should develop with experience.

Relationship with Criterion Measures
A number of studies have correlated performance on simulations with external criterion measures of clinical competence. The most common criterion measure used in these studies was ratings of habitual performance, either from teaching faculty or peers. Norcini and colleagues\(^{(9)}\) reported correlations between peer ratings of practitioners and CASE scores to be 0.30; PMPs correlated 0.28. Skakun and colleagues\(^{(55)}\) reported correlations between individual cases and ratings that averaged essentially zero. Norcini and his co-investigators\(^{(24)}\) reported correlations of 0.3 to 0.4 between PMP scores and ratings from residency programme directors. Helfer and Slater\(^{(1)}\) reported a correlation of 0.4 between clerkship grades and PMP performance. Schumacher\(^{(15)}\) reported correlations of approximately 0.2 with oral examination performance. Taylor and colleagues\(^{(5)}\) reported a correlation of 0.1 between computerised PMP and oral examination performance.

Incremental Validity
Three studies have looked carefully at the improvement in prediction of criterion measures which can be achieved by adding simulations to a test battery of standard MCQs. Using hierarchical multiple regression techniques, Schumacher\(^{(64)}\) found that predictions of group membership improved from a multiple correlation of 0.48 to 0.55 when scores on the INDEX computer simulation were added to internal medicine certifying examination subscores (two multiple choice subtests and a PMP subtest). Using similar regression techniques, Norcini and colleagues\(^{(9)}\) reported that predictions of peer ratings improved from a multiple correlation of 0.32 to 0.34 when PMP scores were added to multiple choice subtests on a recertifying examination. A further improvement to 0.38 was observed when scores on the CASE simulation were added to the predictor set.

Norcini and colleagues\(^{(52)}\) in a large-scale study of the 1980, 1981 and 1982 certifying examinations in internal medicine, observed smaller incremental validities for simulations (PMPs). Predictions of residency director ratings of clinical performance increased from a simple correlation of 0.34 with the multiple choice subtest to a multiple correlation of 0.35 when PMPs were added. Comparable improvement for 1981 and 1982 were 0.35 to 0.37 and 0.36 to 0.38, respectively. Predictions of group membership (two extreme groups composed of highly rated United States Medical Graduates in top residency programmes vs poorly rated Foreign Medical Graduates) improved from 0.65 to 0.653 (1980), 0.69 to 0.72 (1981), and 0.65 to 0.72 (1982). Multiple choice subtests showed much larger incremental validities than PMPs when the order in which predictors were added to the regression equation was reversed. The investigators concluded that MCQs and PMPs measured the same or highly related aspects of clinical competence, but MCQs do so far more efficiently (more measurement information per unit of testing time) because of advantages in broad sampling of content.
Conclusions Regarding Criterion Validity

Overall, the research on the relationship between simulations and multiple choice tests is disappointing. Despite a large number of studies, no conclusions can be drawn because the correlations are attenuated for reliability. More research and better reports of results are needed.

It is clear that performance on simulations is related to group membership and performance on criterion measures. The simulations can distinguish the performance of medical students and residents at different stages of training, although results are mixed concerning the performance of practitioners. Performance on simulations is correlated with ratings of performance in practice to a mild degree, and simulations seem to add a very small amount to predictions over and above traditional multiple choice tests.

Because there are no high quality techniques to assess performance in practice, no “gold standards” by which to judge other measurement techniques, it is unclear how seriously to take the results of the studies that have been conducted. Most of the criterion measures reflect substantially more than the “problem solving ability” simulations are designed to assess. The criterion measures are often subject to the same intercase variability problems that plague simulations. Thus, the reliability and validity of the criterion measures are often unknown, and the probable effect is to depress the magnitude of validity coefficients.

Conclusions Regarding Validity

Methodological problems, the absence of “gold standard” criterion measures, and uninterpretable patterns of results make it difficult to judge the validity of written simulations. There are a number of encouraging studies demonstrating that simulations have the expected relationships with a variety of other indicators of clinical competence. However, there is little evidence that simulations provide unique measurement information not available through traditional MCQ formats in far less testing time. Hence, it is unclear if simulations have any real advantages beyond greater face validity.

DISCUSSION

Summary of Conclusions

The purpose of this paper was to review the research on written and computer based simulations. This review was divided into four major areas: fidelity, scoring, reliability and validity. In each of these areas, the paper synthesised the findings of the better studies and summarised the current state of the art.

Fidelity

Studies of fidelity indicated that simulations provide a reasonably realistic testing vehicle, but that cueing (or lack thereof) does affect performance. In the cued formats, examinees tend to gather more data than they do in uncued formats. Since data
gathering activities with real patients are open-ended, it would seem that uncued simulation formats are more appropriate for assessing these aspects of competence. For other aspects of competence, such as diagnosis and management, cued formats have an advantage because of their efficiency.

Scoring
Studies of scoring indicate that changes in the number of categories into which options are classified and alterations in the weights assigned to those categories make little difference in the resulting scores. Moreover, Aggregate Scoring schemes and Category Scoring schemes produce similar results. This work suggests that the simplest scoring and weighting systems are probably the best, and additional research on these two scoring approaches is unlikely to be productive. Several new scoring techniques have been developed, but these await further research.

Reliability
Studies of reliability indicate that error variance attributable to answer keys is sufficiently large that multiple scorers should always be involved in key development. Variability in examinee performance from case to case is very large, and, consequently, intercase reliability is quite low, unless large numbers of cases are included in a test. Use of shorter cases should permit broader sampling per unit of testing time, thus increasing reliability, perhaps without losing the advantages hypothesised for simulation formats.

Validity
Methodological problems, the absence of a "gold standard" criterion measure and suspect content validity make this a difficult but important area of investigation. Studies of validity have demonstrated that simulations have the expected relationship with a host of criteria. However, it appears that similar assessment information can be obtained with other item formats in less testing time.

Suggestions for Future Research

Methodological Suggestions
The methodological/psychometric quality of much of the research on clinical simulations was strikingly poor. Many studies used small "convenience" samples. Instrumentation and scoring procedures are often described poorly. Reliability coefficients may not be reported, and, when they are reported, the estimation technique is often unspecified, incorrect, or imprecise with the small number of subjects or cases used in the study. Reliability estimation should use the framework provided by generalisability theory, rather than classical test theory and estimation should be based upon case, not item scores; use of the latter yields biased, spuriously high reliability estimates. Observed correlations are usually reported as validity coefficients. These are subject to attenuation due to low reliability, a serious problem. Since validity studies usually require interpretation of patterns of intercorrelations, disattenuated (true), as well as observed, correlations should be reported.
Improved Scoring Procedures
While most of the simple permutations on Category Scoring and Aggregate Scoring have been explored without any clear gains in score reliability or validity, some new directions merit further research. For written simulations, the work of Marshall(19) and Webster(31) seems promising. For computer-based simulations, scoring methods developed in conjunction with the CBX simulation deserve follow-up.(10)

Mini-PMPs
The major psychometric problem with simulations is poor reliability per unit of testing time. It may be possible to make simulations more efficient in use of testing time by shortening individual cases. One- or two-scene PMPs may well provide almost as much measurement information as longer cases. These “mini-PMPs” may provide a nice blend of the face validity of PMPs and the efficient sampling characteristics of MCQs: by lowering the fidelity of the simulations, their psychometric characteristics may improve.

Use of Item Response Theory
The direct consequence of poor reliability is a large standard error of measurement, or equivalently, poor precision. Item response theory (IRT) has proven useful in improving the precision of MCQ-based examinations.(68) It provides analytic tools to identify questions which yield maximal measurement information in a particular ability range (e.g. near the pass/fail score in a licensure/certification examination). Similar techniques may be useful in improving the precision of simulation-based tests, by selecting the most discriminating cases for a particular testing purpose. This application of IRT has not been explored to date, though appropriate scaling models appear to exist.

Sequential Test Administration
In use of simulations to date, the same set of cases is administered to all examinees. Normally, this is unnecessary for those examinees who are fairly distant from the pass/fail score: it is clear that they will pass or fail after a few cases. A sequential testing strategy could be adopted to take advantage of this fact. Examinees who will clearly pass or fail could be identified using a relatively brief set of PMPs; they could then be excused from the examination. The test could then be continued for the remaining examinees to achieve greater precision. This approach would shorten average test length across examinees, at the same time improving precision for those examinees where it is needed. IRT-based techniques (adaptive testing models) might be useful here as well, though they are untried.

Criterion-Referenced Testing
Virtually all research on PMPs, at least implicitly, has been within a norm-referenced testing framework: score interpretation is based upon an examinee’s performance relative to other examinees. Criterion-referenced test scores, on the other hand, allow performance to be interpreted relative to some domain of behaviour or knowledge.
Intuitively, because simulations require examinees to perform real-world tasks under realistic circumstances, it would be possible to interpret scores within a criterion-referenced testing framework. Preliminary work in this area has been encouraging, but more research is needed.

Recommendations for Use of Clinical Simulations

While clinical simulations pose some interesting research challenges, until research efforts improve their psychometric characteristics, they should probably not be used. The scores resulting from short tests, involving only a few simulations, are not meaningful because of poor reliability. Use of clinical simulations as one component of a test battery to improve "face validity" is, perhaps, justifiable, as long as separate scores are not reported for the simulation component, and little weight is given to it in calculation of the composite score for the battery. Longer (one- to two-day) simulation-based tests may be useful from a reliability perspective, but research to date suggests that clinically oriented multiple choice questions would yield a similar rank-ordering of examinees. In particular, it seems difficult to justify inclusion of written simulations in tests used for licensure or certification, where efficient use of testing time is crucial, and important decisions depend upon precise, reliable measurement.

REFERENCES


Further details from:
David B. Swanson, PhD,
American Board of Internal Medicine,
3624 Market Street,
Philadelphia, Pennsylvania,
USA 19104.

David Swanson is a computer scientist and psychometrician. He is Assistant Director of the American Board of Internal Medicine – the body responsible in the USA for accreditation of doctors specialising in medicine (as opposed to surgery, family practice).